**Life Expectancy Analysis: A Data Science Approach**

**Subtitle**: "Analyzing Factors Affecting Life Expectancy in Japan Using Data Science"

**Date**: January 2025

**Name**: Khalil Mosbah, Data Scientist

**Executive Summary:**

**This project analyzes key factors affecting life expectancy in Japan. Using Linear regression, OLS regression and SHAP analysis we found that *education*, *healthcare access*, and *park availability* in a prefecture are the most significant predictors of life expectancy in our data set.**

**Data & Methodology:**

* **Data: Kaggle Data Set by Gianina-Maria Petrașcu**

<https://www.kaggle.com/datasets/gianinamariapetrascu/japan-life-expectancy>

* **Methodology:**
  + **EDA Observations & Feature Engineeing:**

**\*Worst vs best life Expectancy: Aomori 82.8 & Shiga 85.5 years**

**\*Correlation Heatmap:**

**Strong (+) : ‘University\_%’ & ‘Junior\_col\_%’**

**Moderate (-): ‘Physician\_100kP’, ‘Park\_land\_%’ & ‘Salary’**

**(-): ‘Elementay\_school\_%’ & ‘Ambulances’**

**\*VIF Analysis: Severe multicolinearity ‘University\_%’ & ‘Salary’**

**\*Normalization: Applied to all numeric values to reduce skewness**

**\*PCA: Combined ‘Salary’ & ‘University\_%’ into a new feature called ‘Socioeconomic\_index’ to avoid multicolinearity**

* + **Model Evaluation & Feature Importance:**

**\*Linear Regression Model trained on the features ‘Physician\_100kP’, ‘Junior\_college\_%’, ‘Socioeconomic\_index’, ‘Ambulances\_100kP’, & Park\_Land\_% gave the best performance:**

**R² = 0.37 & RMSE = 0.34**

**\*SHAP Analysis:**

**Key Predictors: Physician\_100kP’ & ‘Junior\_col\_%’**

**Moderate Impact: ‘Park\_Land\_%’, ‘Socioeconomic\_index’ & ‘Ambulances\_100kP’**

* + **Model Experiments:**

**\*OLS Regression: Education, healthcare access and Park access are statistically significant predictors of life expectancy while socioeconomic index and Ambulances\_100kP were insignificant.**

* + **Insights:**

**\*Education: 1% Jun\_colege increase -> LifeE. increase 1.2y**

**\*Healthcare access: Additional physician/100kP ->LifeE. Increase 0.7y**

**\*Access to Parks: 1% increase park land-> LifeE. increase 0.56y**

**Recommendations:**

* **Invest in education and healthcare: These are the primary drivers of life expectancy.**
* **Increase park availability: More green spaces could significantly improve public health.**
* **Policy implications: Policymakers should focus on education, healthcare, and urban planning.**

**Visual Elements:**

**Correlation Heatmap**  **Model prediction:**

A screenshot of a graph

Description automatically generatedA graph with blue dots

Description automatically generated

**SHAP Feature Visualization OLS Regression Results**

A graph of a number of dots

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated